

scale

2023 AI Readiness Report

The word "zeitgeist" is rendered in a large, white, serif font. The letters are semi-transparent, allowing a vibrant rainbow prism effect to be seen through them. This effect is most prominent on the left side, where a large sphere and a smaller cylinder are visible, and on the right side, where a faceted crystal structure is shown. The background is a solid black, which makes the white text and the colorful prism effects stand out sharply.

zeitgeist

Table of Contents

AI Year in Review	04		
AI Adoption Trends	16		
Top Use Cases By Industry	22	ML Lifecycle	30
Insurance	24	Working with Foundation Models	31
Retail & eCommerce	25	Data Challenges	33
Financial Services	26	Data Best Practices	35
Logistics & Supply Chain	28	Model Evaluation	37
		Conclusion	40
		About Scale	42
		Methodology	44

Introduction

At Scale, our mission is to accelerate the development of AI applications. That's why we are excited to introduce the 2nd edition of Scale Zeitgeist: AI Readiness Report, a survey of more than 1,600 executives and ML practitioners to uncover what's working, what's not, and the best practices for organizations to deploy AI for real business impact.

Over the past several months, hype about Generative AI has flooded popular discourse. Seemingly overnight, every CEO began adding Generative AI to their business strategy. Many influencers and leaders have opposing views of Generative AI: some are worried about highly competent AI systems causing mass job displacement, while others find these models are not enterprise-ready due to privacy and security concerns. We are here to get past the hype and help you separate the signal from the noise on what it really takes to adopt Generative AI to help you solve your most important business challenges.

Despite the widespread interest in Generative AI, with 72% of survey respondents planning to increase their AI investments this year, we found that large enterprises, small companies, and governments are still uncertain about how to adopt this technology. Executives are quickly realizing that you can't take these models off the shelf and expect to get a unique business advantage – you need to train them to perform for your specific business needs, with your proprietary data.

We designed this report to explore the real impact of Generative AI, how every industry can adopt AI to accelerate innovation, and the best practices companies can follow to maximize their investment in the technology. The results of this report show that with the right foundation in place for implementing AI, employee productivity increases, customer experience improves, and revenue and profitability grow.

Our goal is to continue to shed light on the realities of what it takes to unlock AI for every business. The AI landscape is moving fast, but we hope that this research helps demystify the changes that are underway and sheds light on what businesses need to take full advantage of this moment.



The next 2-3 years of AI are definitely going to define the coming 2-3 decades of the world.

For those in technology: you live a lifetime for a moment like this—don't waste it.

There are decades where nothing happens, and weeks where decades happen.

—Alexandr Wang
FOUNDER & CEO, SCALE

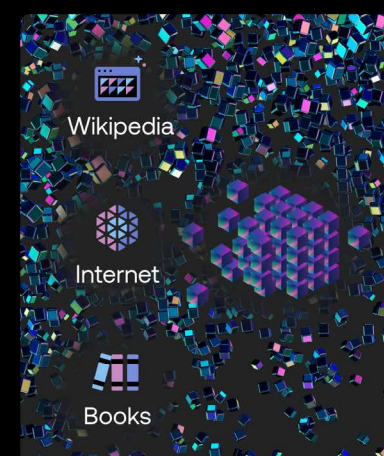


AI Year in Review

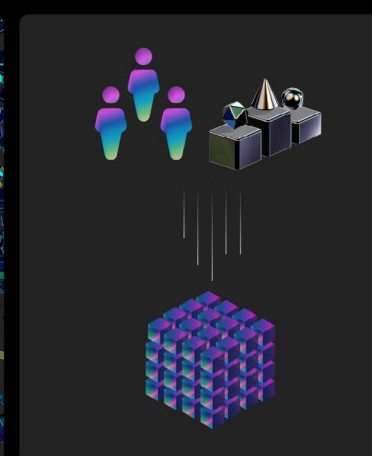
Generative AI takes a giant leap

Giant leaps in the capabilities of Generative AI defined 2022. In mid-2022, the image diffusion models Dall-E 2 (OpenAI) and Stable Diffusion (Stability AI) captured headlines. Many new startups rushed to build their businesses on top of these powerful image-creation tools. In late November, OpenAI released the large language model GPT-3.5 and the chat interface ChatGPT which quickly became one of the most impactful technology launches ever. Key to the impact of this model was the use of Reinforcement Learning with Human Feedback (RLHF), a technique to align model performance with human intent.

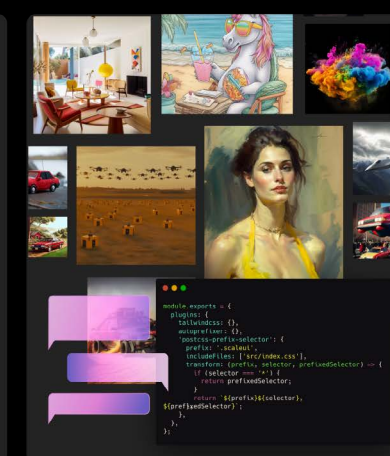
What is Generative AI?



Vast amounts of data are collected from a variety of sources, primarily data from the internet. A model is trained on GPUs, producing a base model that is highly capable, but not aligned to human preferences.



The base model is then fine-tuned via Reinforcement Learning from Human Feedback (RLHF) to align them more closely to human preferences and in the case of LLMs interact in a conversational or “chat” format.



Generative models are then able to produce unlimited and infinitely creative imagery, engage in conversations with users, summarize documents, and write code.

In March 2023, OpenAI launched GPT-4, the most capable large language model ever created. Also aligned with RLHF, GPT-4 impressively passed many exams designed to test human capability in the 90th percentile or higher, including several AP exams and the bar exam. In December 2022, Anthropic announced a closed beta of Claude, an LLM with chat capabilities similar to ChatGPT, and an approach to human alignment based on “[constitutional AI](#),” where human oversight is provided through a list of rules or principles. In March 2023, Google released Bard, its conversational AI based on Google’s LaMDA model. Several other companies are also building large language models, including AI21 labs, Carper AI, Stability AI, and Cohere. Domain-specific models will

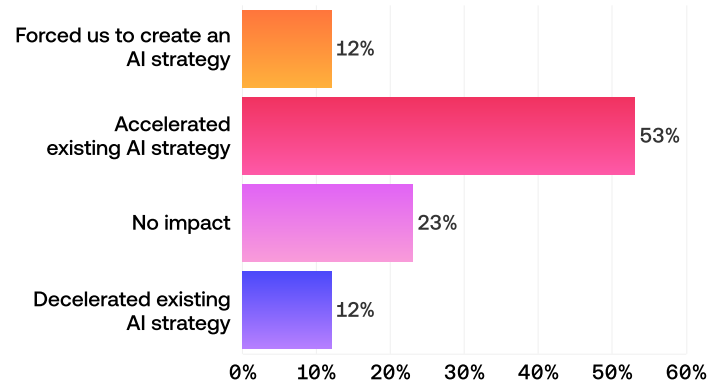
also be developed over the coming year, like BloombergGPT, a model purpose-built for financial use cases. A steadily growing number of these models will be trained and released over the next year.

Generative models are being integrated into Google Workspace and Microsoft Office, enabling massive productivity gains for business users. These tools help you write the first draft of a document, generate complete presentations from a single prompt, or automatically analyze and visualize financial data in a spreadsheet. Enterprise software platforms like Salesforce are enabling analysts and executives to gain new perspectives on data using Generative AI.

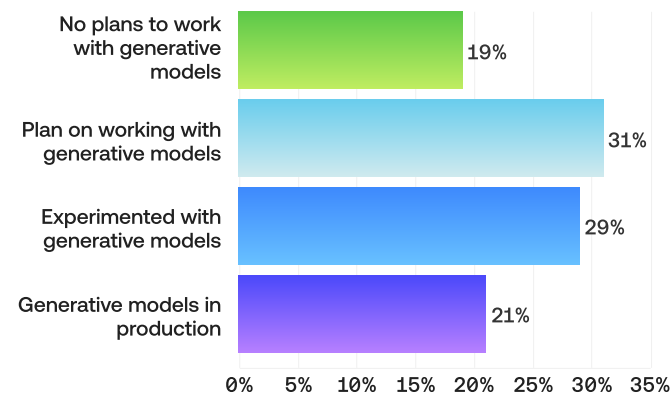
The significantly improved capabilities of generative models in 2022 enormously impacted businesses' AI strategies, with 65% either accelerating their existing strategies or creating an AI strategy for the first time.

While most respondents (60%) are experimenting with generative models or plan on working with them in the next year, only 21% have these models in production. Companies see the potential of generative models to improve their business, but getting them into production is challenging. To unlock the power of their data and take full advantage of these models, companies need machine learning expertise, fine-tuning infrastructure, and the resources to perform RLHF at scale.

To what extent did advances in generative models in 2022 impact your AI strategy?



Which of the following describes how your company works with generative models?



65% either accelerated their existing strategies or created an AI strategy for the first time

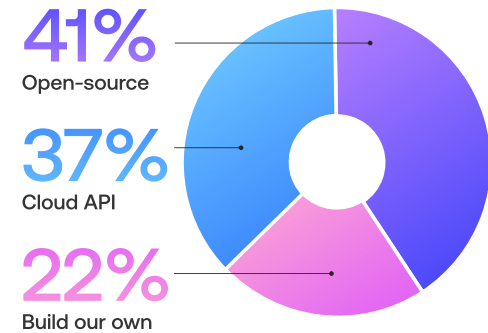
“I think that for this decade, what we’re really going to see is these tools proliferating. They’re going to be everywhere. They’re going to be baked into every company. I think it’s kind of like the internet transition. If you’re a company, what’s your internet strategy? And here we are today, where we don’t even talk about internet strategies. It’s just so integral to every business. It’s not a separate part of your business that you can pick or choose whether you’re going to have it. And I think that AI is going to be much the same.”

— GREG BROCKMAN,
PRESIDENT & CO-FOUNDER, OPENAI

(TransformX October 2022)

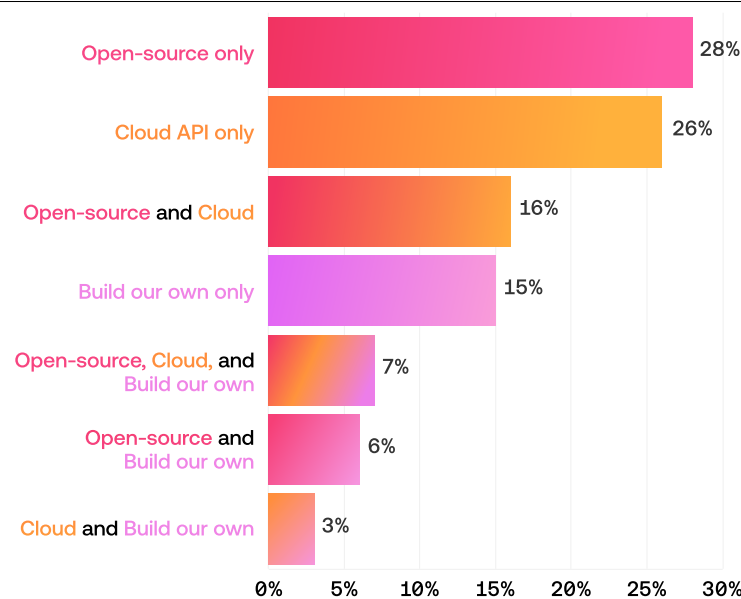
Most companies don't have the necessary resources or mandate to create their own generative models, so they must rely on third parties. Of those companies that plan on working with generative models, the vast majority are looking to leverage open-source generative models (41%) or Cloud API generative models (37%), while very few are looking to build their own generative models (22%).

How do you work with generative models?



Furthermore, 28% are exclusively using open-source models, while 26% use cloud APIs (commercially available models such as Anthropic's Claude, OpenAI's GPT-4, and Cohere's Command), and only 15% are exclusively building their own.

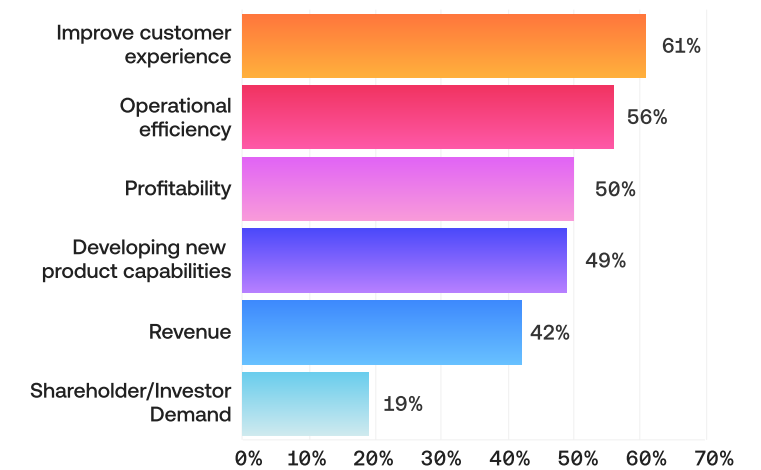
How do you work with generative models?



There are multiple factors that enterprises must weigh when deciding on their Generative AI infrastructure, including their in-house machine learning expertise, budget, security requirements, and need for domain-specific capabilities. Leveraging a cloud provider is the easiest and fastest path to obtain generative capabilities, but it comes with higher security risk, less control over the underlying models, lower quality performance at domain-specific tasks, and can be expensive. Open-source models provide more control and are cheaper, but they require more in-house expertise to deploy and fine-tune. Companies looking to build their own generative models benefit from greater control but incur greater costs from data collection, compute, and hiring machine learning experts to train and deploy them.

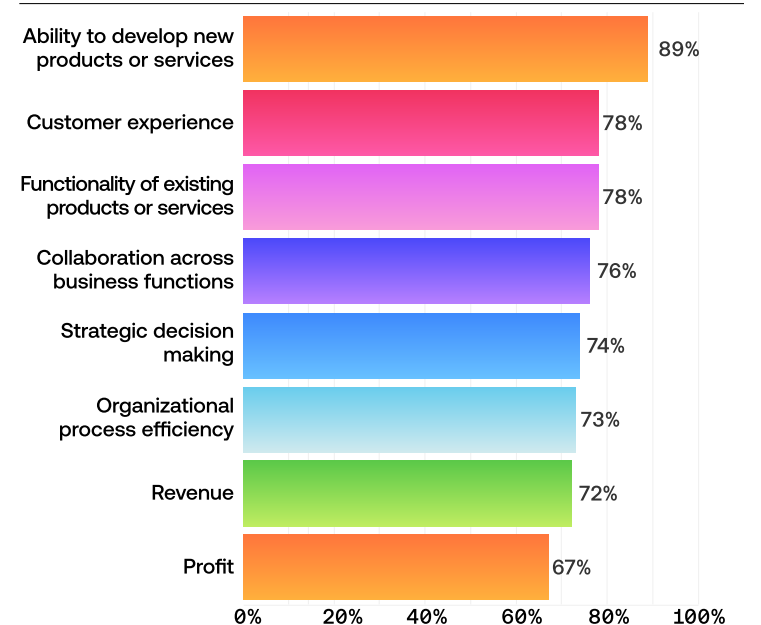
61% of companies are looking to AI to help enhance the customer experience, 56% to improve operational efficiency, and 50% to increase profitability. Focusing on customer-centricity benefits organizations immensely, with improved customer goodwill in the short term and greater profitability in the long term.

Which goals describe why you are adopting AI at your organization?



89% of companies adopting AI benefit from the ability to develop new products or services, 78% from enhanced customer experiences, and 76% from better collaboration across business functions. These companies also see improved organizational process efficiency and profitability. Despite the positive outcomes for AI adopters, even greater outcomes are possible as companies accelerate their AI strategies and increase their investments in AI.

What outcomes have you seen from AI adoption?



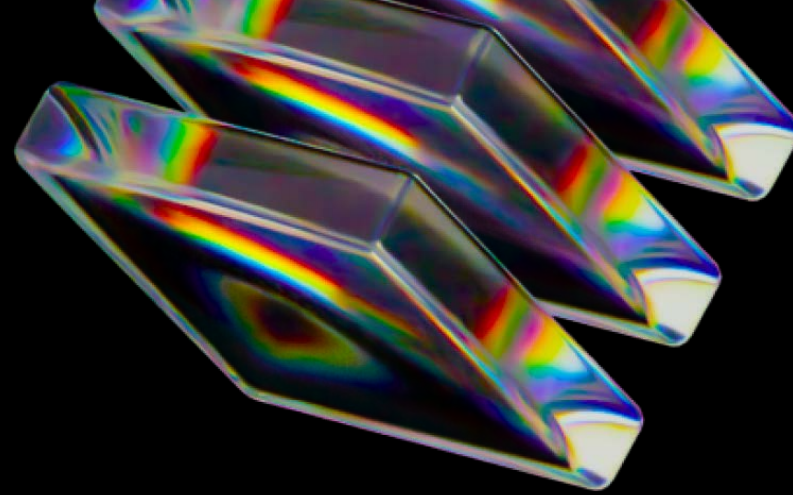
An organization's goals also shape the effectiveness of its AI implementation. Those companies that list shareholder/investor demand as the top goal for AI adoption also show the poorest results for customer experience, revenue, and profitability. To ensure success with an AI implementation, organizations must avoid implementing AI for the sake of implementing AI, but instead, ensure the goals of an AI implementation are aligned with company priorities and that AI is a good solution for a given problem.

A New Era

Generative models are already transforming how we create art, understand our world, and conduct business.

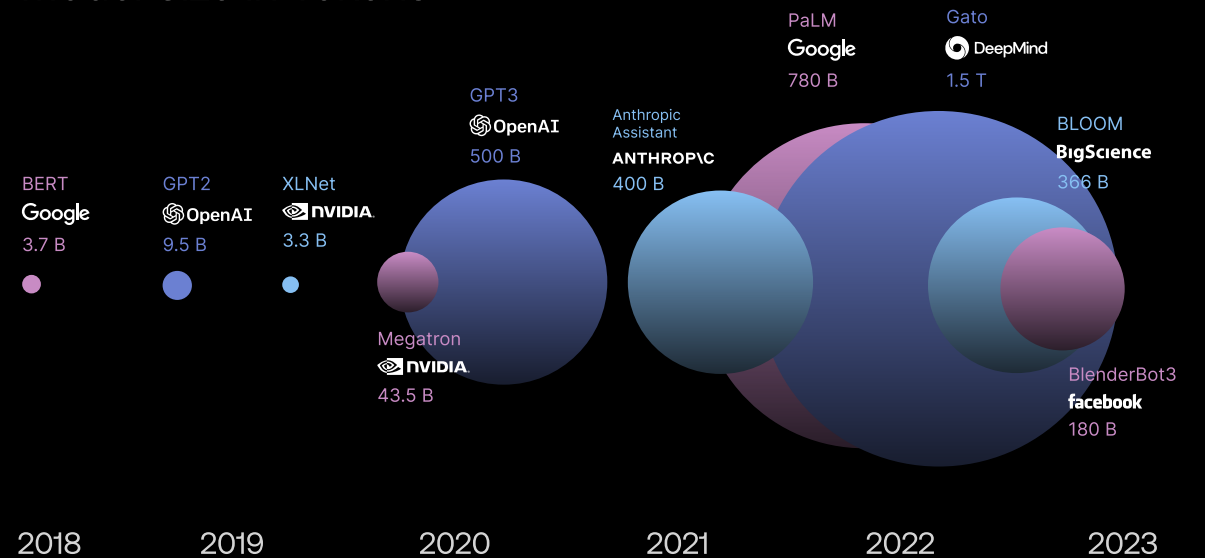
Large language models help us write content such as blogs, emails, or ad copy more quickly and creatively. They summarize long-form content so that we can quickly understand the most critical information from reports and news articles. Diffusion models streamline marketing workflows, enabling marketers to generate unlimited and infinitely creative product imagery. Developers use LLMs to write code more efficiently and help them quickly identify and fix bugs. Advanced chatbots enable businesses to improve their customer service at a lower cost. Finally, organizations are unlocking the power of their knowledge bases by customizing LLMs with their proprietary data to perform better on tasks unique to their business.

We will now look at a few key terms and trends essential to understanding this new era of Generative AI.



Models Are Increasing in Size

Model Size in Tokens



Over time, generative models have become more capable as they've increased in size. Model size is typically determined by its training dataset size measured in tokens (parts of words) or by its number of parameters (the number of values the model can change as it learns).

- [BERT \(2018\)](#) was 3.7B tokens and 240 million parameters.
- [GPT-2 \(2019\)](#) was 9.5B tokens and 1.5 billion parameters.
- [GPT-3 \(2020\)](#) has 499B tokens and 175 billion parameters.
- [PaLM \(2022\)](#) was 780B tokens and 540 billion parameters.

As these models scale in size, they become increasingly capable, providing more incentive for companies to build applications on top. Generative models are now more widely available as many large model developers provide APIs or make them open-source, and companies are quickly adopting these large models to

their specific business use cases.

Generative models are trained on a large amount of internet data, making them competent generalists. These models can write poetry, solve logic puzzles, and identify bugs in code. While generative models are great generalists, they are poor specialists when solving problems outside of their data distribution. Since a significant portion of data is proprietary to individual organizations, base large language models are not well adapted to these specific domains. To improve performance on the specific tasks of, say, an insurance company, an eCommerce company, or a logistics company, these models must be fine-tuned and aligned to excel at those particular tasks and provide responses that are useful to customers and employees.

Reinforcement Learning from Human Feedback (RLHF)

Though Reinforcement Learning from Human Feedback (RLHF) is not new to the research community, in 2022, it catapulted in popularity as it was a critical ingredient in the success of ChatGPT.

Instead of attempting to write a loss function with which to train a model, RLHF involves soliciting feedback from human users and training a reward model on that feedback. This human-defined reward model is then used to train a base model. This also allows training on much more data since the human feedback is mimicked by the reward model, so the dataset size is now only constrained by how many prompts you can create.

RLHF tuning results in models [better aligned to human preferences](#), producing more detailed and factual responses.

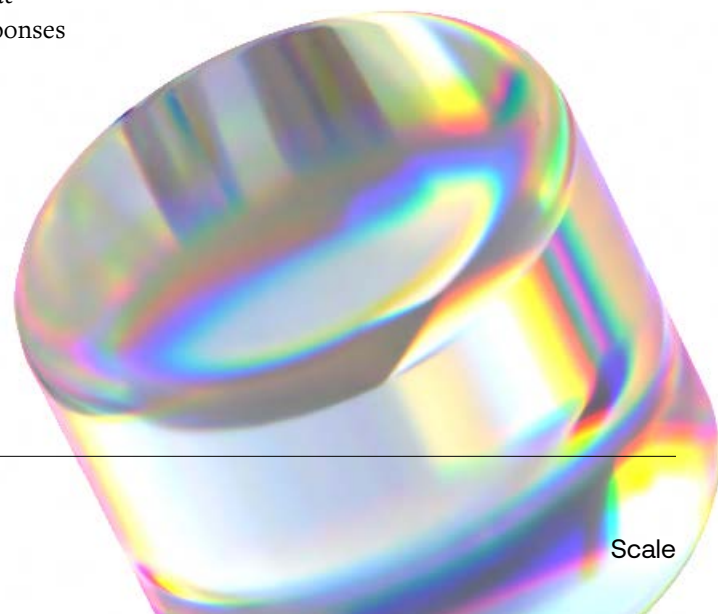
RLHF also defines the “personality” and “mood” of the model, making it more helpful, friendly, and factual than the base model would be otherwise. This means we get responses from the model that feel more human and less like talking to a machine. RLHF is a critical component to the success of recent LLMs and is also critical to ensuring that enterprises using Generative AI get model responses that align with their policies and brands.

ChatGPT

ChatGPT is a large language model that has been tuned specifically for the task of conversational text generation. ChatGPT was trained with RLHF and data in dialogue formats to enable it to act as a conversational chatbot.

ChatGPT quickly became one of the most impactful product launches of all time, reaching 1 million users in just five days and currently sits at over 100 million users.

ChatGPT was initially launched with GPT-3.5 but now also includes GPT-4 for ChatGPT plus subscribers. These models are highly capable of question answering, content generation, and summarization. While these models provide more robust, informative, and creative responses than their predecessors, the real breakthrough for adoption was their ability to hold conversations with humans. The ability to interact with the models in an intuitive way increased the accessibility of the models so that anyone can use them.



Prompt Engineer

2022 also saw the proliferation of a new role for machine learning teams, the “prompt engineer.” While generative models are competent at generating the desired output for business use cases, the right prompt is required to optimize the model’s effectiveness. Prompt engineers carefully craft natural language inputs to get more consistent and reliable responses from models so that the model outputs can then be used in business applications. Instead of writing an SQL query, these engineers craft finely honed natural language prompts.

For example, when integrating applications with LLMs, the varied responses from the model can break the integration if they are not formatted properly. Say you are creating an application dealing with financial data. A user’s input may be related to the Q4 earnings of a particular company (see graphic, right).

Prompt engineers help model responses to solve business challenges with greater accuracy, efficiency, and quality. They also ensure that responses align with an organization’s brand guidelines and voice. They are also critical at finding vulnerabilities in models by using prompt injection techniques and helping to resolve those vulnerabilities before an employee or customer does. This role is highly valuable in ensuring organizations’ successful implementation of generative models.

USER INPUT:

What was the 4th quarter 2022 revenue of Generic Corp.?

RESPONSE:

\$76,028 million

This response provides an accurate answer but does not include all of the detail that your application needs and is in a format that your application will not understand and will throw an error. So instead, a formatting prompt is applied to all user input to format the response properly.

FORMATTING PROMPT:

You are a financial services bot, and you should format responses to requests for financial information using the following template:

```
{
  'company':'Company Name',
  'ticker':'Ticker',
  'period':'period',
  'revenue':'revenue',
  'units':'unit',
  'currency':'currency'
}
```

USER INPUT:

What was the 4th quarter 2022 revenue of Generic Corp.?

RESPONSE:

Here is the information you requested:

```
{
  'company':'Generic Corp',
  'ticker':'GENERIC',
  'period':'Q4 2022',
  'revenue':'$76,028',
  'units':'million',
  'currency':'USD'
}
```


“One of the ways that I describe Stable Diffusion is as a generative search engine. You don’t need to use image search anymore cause you can just make the image. By putting this in a pipeline at the right place and having human-in-the-loop interactions like the work that you do at Scale AI, understanding how humans do that at scale, and having these engines, it will allow us to have even better experiences that understand what we want.”

**—EMAD MOSTAQUE,
FOUNDER & CEO, STABILITY AI**

(TransformX October 2022)

What to Expect in 2023

Increased investment in AI

As Generative AI is now more capable and widely available, companies are quickly incorporating it into their operations. 72% of companies will significantly increase their investment in AI each year for the next three years.

Increasing capabilities of generative models

Many organizations are now building their own large generative models. These models are being incorporated into search engines and paired with other tools, including internal knowledge bases, to become powerful business tools. These models will also become multimodal, meaning that they will be able to consume and generate text, images, and video, making them even more useful than they are today. You can upload a product catalog with both images and text to a multimodal model, and it will recognize specific products, write product descriptions, fill in missing attributes, and generate new images to enrich your product catalog automatically.

Widespread accessibility of generative models

Much like the cloud, widely accessible generative models represent a paradigm shift for companies. Incorporating this type of AI will quickly become the status quo, and those who are slow to adopt will be left behind.

Proprietary data will unlock the power of generative models

On their own, base generative models are valuable tools. Paired with a business’s proprietary data, they become strong differentiators, improving the customer experience, product development, and profitability.

“I’m most excited about...the ability for our models to start using tools out in the world. Giving them access to knowledge bases, giving them access to search engines like Google or Bing, and augmenting them with that knowledge as a resource so that instead of them having to memorize all of these facts, they can make reference to a live, updated, knowledge base. I think that’s going to be super impactful.”

**—AIDAN GOMEZ,
CEO, COHERE**

(TransformX October 2022)

AI Adoption

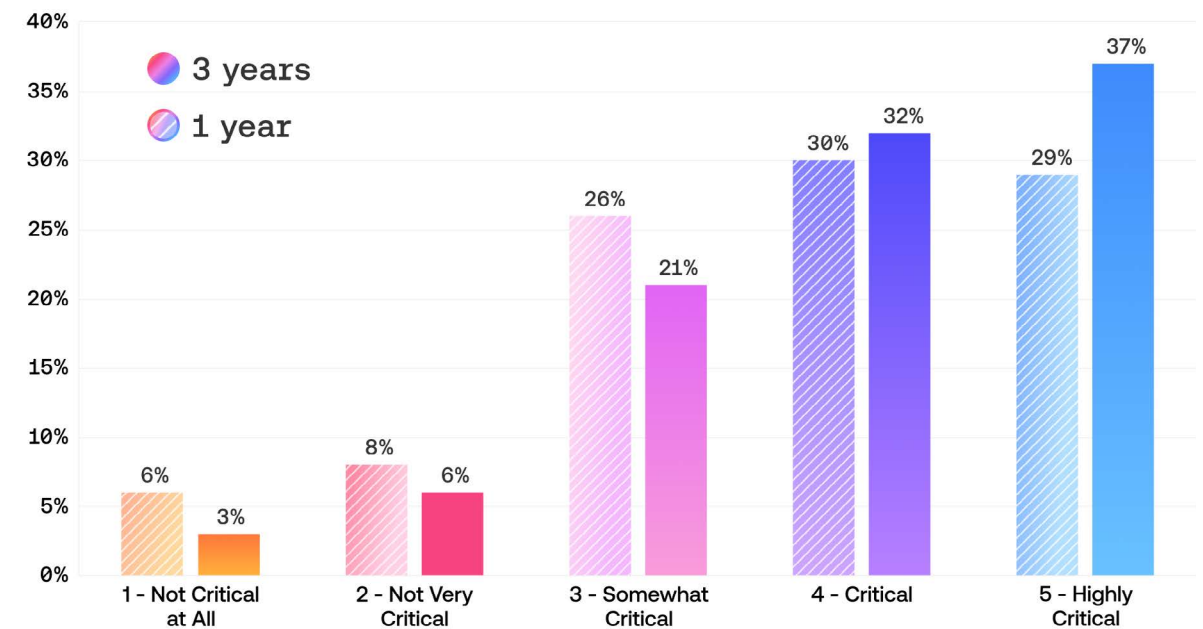


Adoption Trends

Business leaders have identified that AI is critical to the future of their companies and are looking to adopt it as quickly and with as much impact as possible. We examine this trend and provide insights on best practices.

72% of companies are looking to increase their investment in AI each year for the next three years

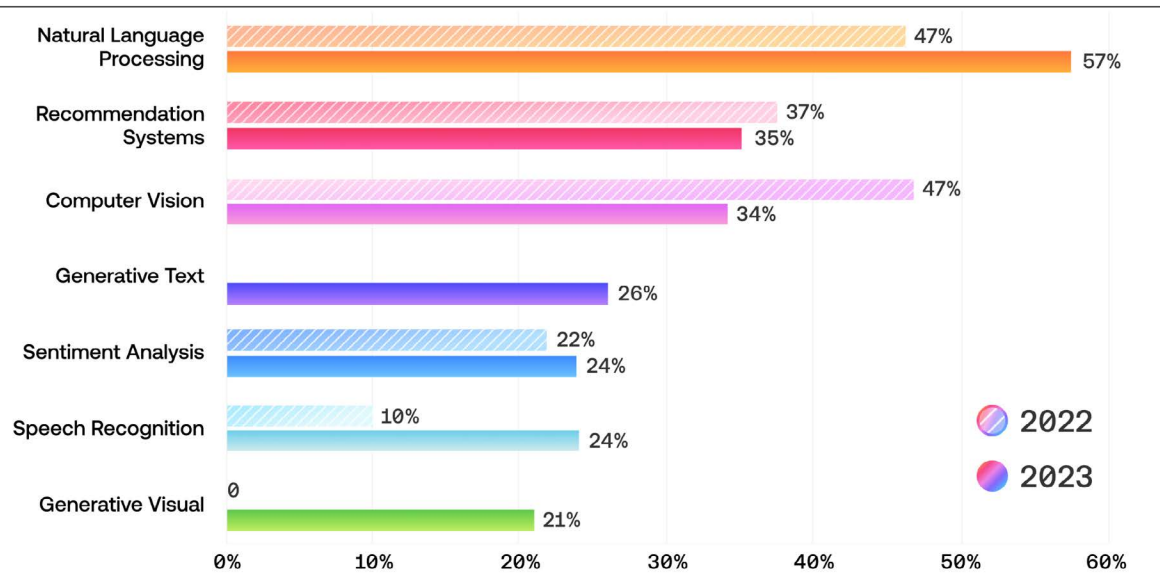
How critical is AI to your business in the next 1 - 3 years?



59% of companies view AI as critical or highly critical to their business in the next year, and 69% in the next three years. The increasing capabilities and availability of Generative AI will accelerate AI adoption.

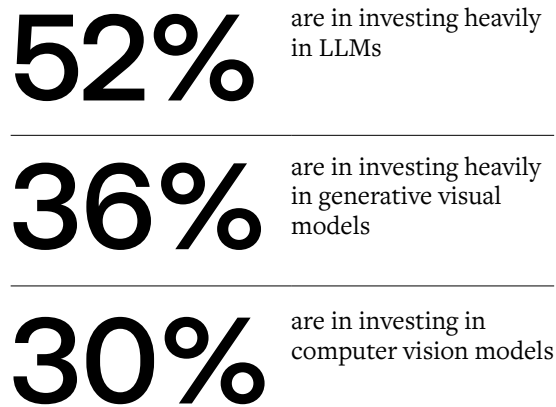
As companies view AI as more critical to the future success of their business, they are increasing AI investments over the next three years. 72% of companies plan to increase their investment in AI each year for the next three years.

What type of ML systems do you work on?



Companies are spending less time and effort on traditional computer vision applications and instead focusing on LLMs and Generative AI. Of companies making significant investments in AI, 52% are investing heavily in LLMs, 36% in generative visual models, and 30% in computer vision applications. With the recently popularized capabilities of LLMs, companies have rapidly shifted their AI strategies to harness the power of Generative AI.

Of companies making significant investments in AI:

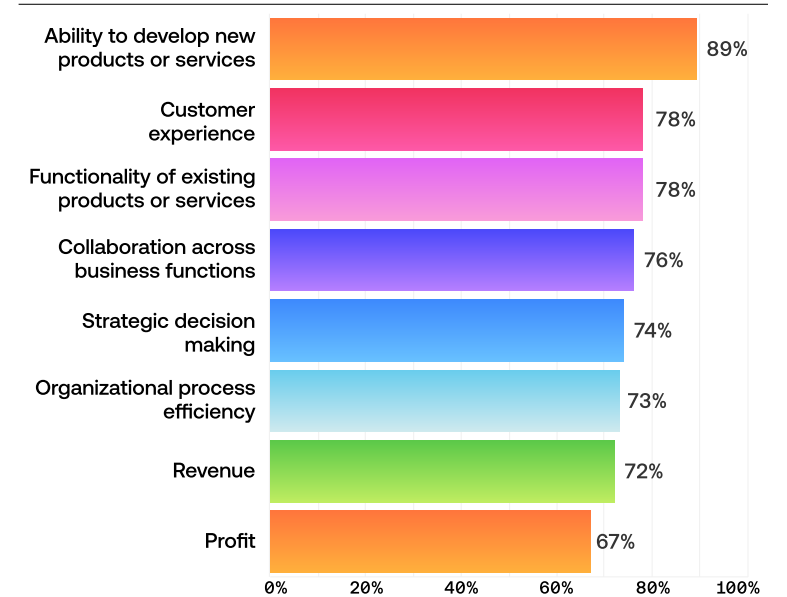


What outcomes are achieved by companies that adopt AI?

As mentioned previously, companies adopting AI are seeing positive outcomes from improved customer experiences, the ability to develop new products or services and improve existing products, and improved collaboration across business functions.

Across the board, companies adopting AI are achieving positive outcomes in almost every category. Like any technology program, the success of AI programs depends on the ability to implement AI and align implementation efforts with measurable organizational goals.

What outcomes have you seen from AI adoption?



“I really believe that we are at a transformative moment today where ML is moving at an incredible speed and problems that were thought to be too complex to solve with computers a few years ago, are now being solved by applying machine learning. So we have this great opportunity. If machine learning becomes more accessible, the world will move faster, our economy will move faster, science will move faster.”

**— FRANCOIS CHOLLET,
AI ENGINEER AND RESEARCHER, GOOGLE**

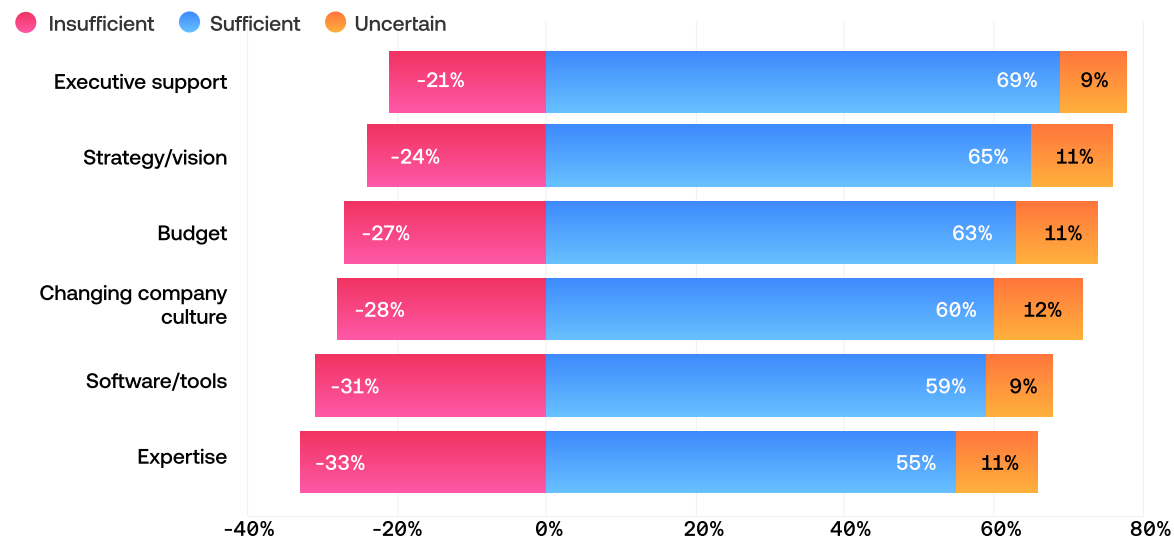
(TransformX October 2022)

Which resources do companies feel they have enough of in order to successfully implement their AI strategies? Which resources are they lacking?

Companies that view AI as critical to their business indicate they have the executive support, strategy and vision, and budget they need to succeed in implementing their company's AI strategy. However, these companies generally lack the necessary expertise, software, and tools required to achieve success.

While leaders have identified the need to adopt AI, the execution of these strategies is difficult, nuanced, and heavily dependent on expertise. The field is moving so quickly that it is difficult to keep up with the pace of advancement. Highly talented people with expertise in Generative AI are simply not available to most organizations. Similarly, selecting, standardizing, and updating the software and tools associated with Generative AI, MLOps, and even DevOps is challenging for companies without dedicated teams to keep up with these changes as the requisite tech stacks are constantly evolving.

Which resources do companies feel they have enough of in order to successfully implement their AI strategies?



“As a product of this shortage in AI talent, most businesses are missing out on a huge opportunity to integrate this tech into products and into their developer’s workflows. Consumers are missing out on products that have more magical, intuitive, and smart experiences. The start of the fix comes with the product folks making the decision about what’s prioritized, looking at what they can do, understanding where the technology is today, and where they could insert it, and then building it. I think we need to start making AI a standard piece of every single product. I don’t think consumers are going to tolerate dumb products anymore. We need to make them much, much smarter.”

**— AIDAN GOMEZ,
CEO, COHERE**

(TransformX October 2022)

AI Adoption by Industry

Every industry is looking to increase its AI budgets over the next three years. Those that top the list are:



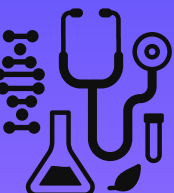
80%
Insurance




79%
Logistics & Supply Chain



77%
Financial Services



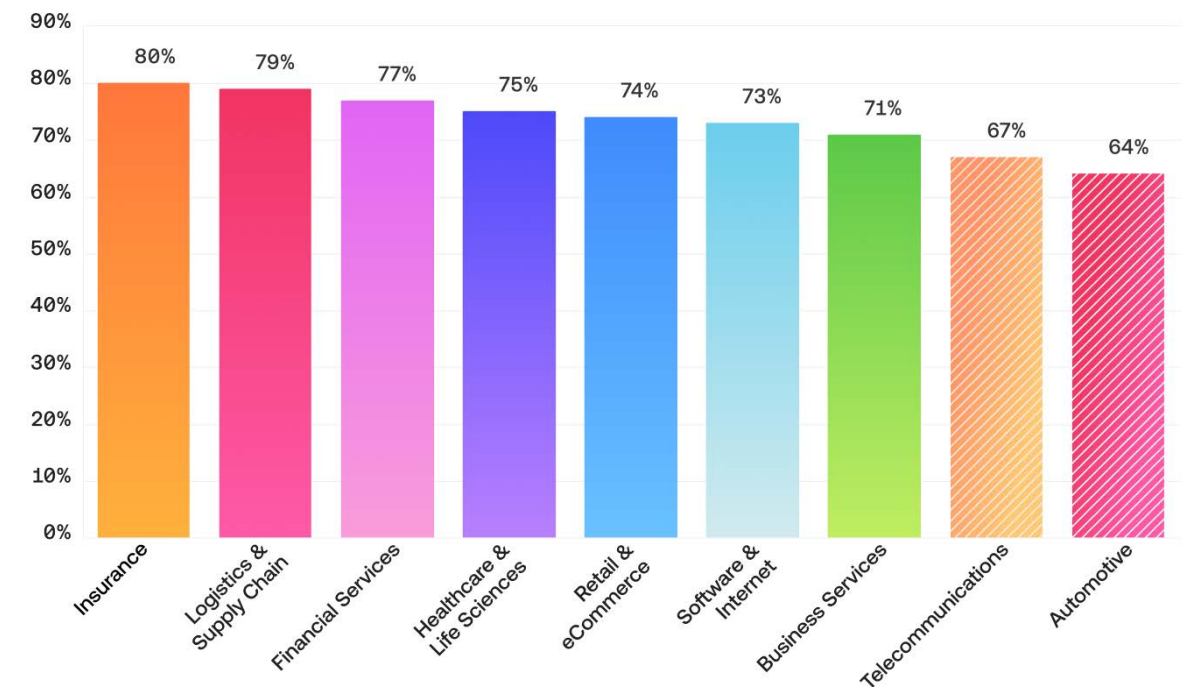
75%
Healthcare & Life Sciences



74%
Retail & eCommerce

TOP USE CASES BY INDUSTRY

Does your company plan on increasing its AI budget in each of the next 3 years?



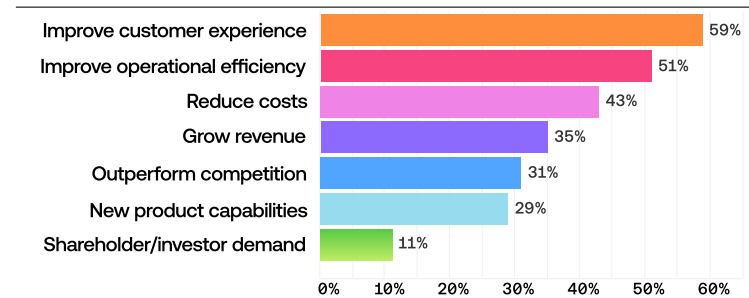
While all industries are increasing their AI budgets, each industry has unique use cases. These range from insurance companies looking to reduce claim processing times to eCommerce companies implementing customer service chatbots. We examine how a few key industries are adopting AI in 2023.



Insurance

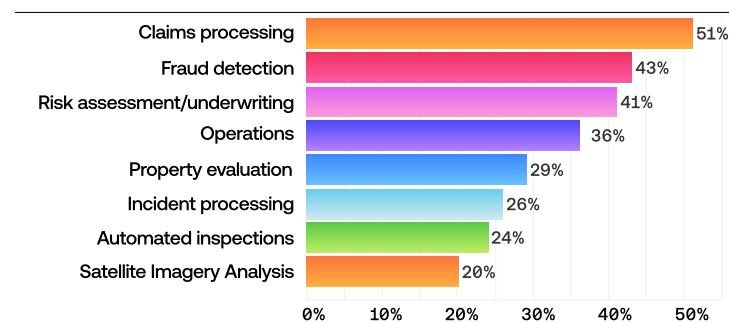
Insurance companies look to AI to help them improve customer experience and improve operational efficiency.

Which goals describe why you are adopting AI at your organization?



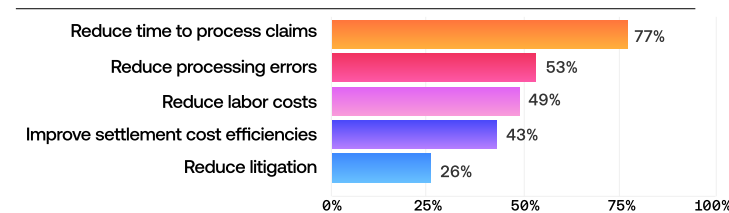
To help achieve these goals, insurance companies are looking to adopt AI to improve claims processing, fraud detection, and risk assessment/underwriting.

Which of the following will your company use AI to address?



For claims processing in particular, insurance companies believe that AI can help reduce time to process claims and reduce processing errors which will result in a better experience for their customers and improved operational efficiency.

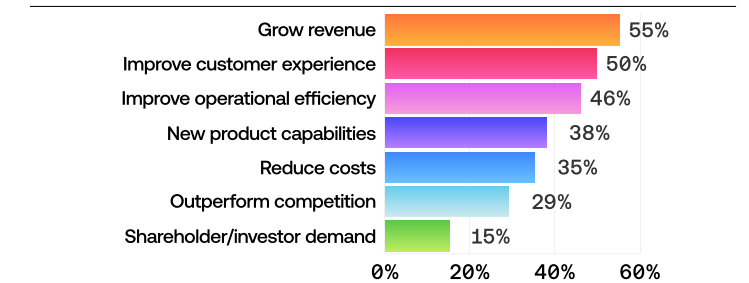
What outcomes are you expecting to achieve from implementing AI for claims processing?



Retail and eCommerce

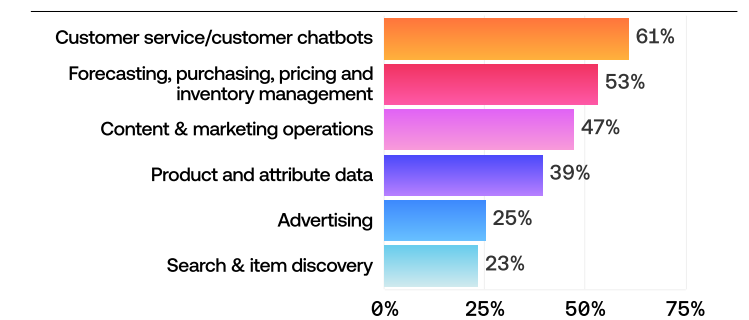
Retail and eCommerce companies look to AI to help them grow revenue, improve the customer experience, and increase operational efficiency.

Which goals describe why you are adopting AI at your organization?



To help achieve these goals, retail and eCommerce companies are adopting AI to improve the customer experience with more capable chatbots. They also want to improve operational efficiency with more productive content and marketing operations built on AI-generated product imagery and descriptions. These companies are also enhancing their operational efficiency with better forecasting, purchasing, pricing, and inventory management. Retail and eCommerce companies are not focusing on adopting AI to directly grow revenue but are indirectly looking to influence revenue growth with increased operational efficiency and an improved customer experience.

Which of the following will your company use AI to address?





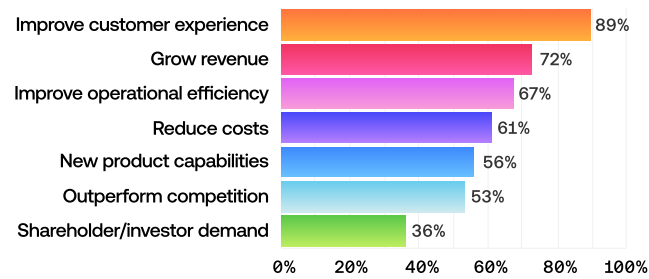
Financial Services

Financial services companies look to AI to help them enhance the customer experience, grow revenue, and increase operational efficiency.

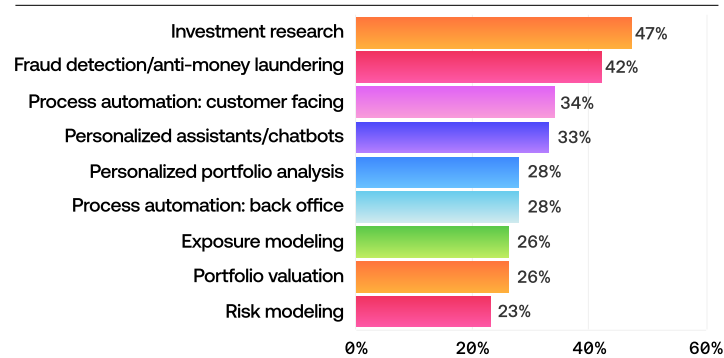
To help achieve these goals, financial services companies are looking to adopt AI to improve investment research, fraud detection, customer-facing process automation, and to power personalized chatbots.

For investment research in particular, financial services companies are applying AI to summarize content, detect trends, and classify topics to improve investment decisions, resulting in increased revenue and improved operational efficiency. By improving their investment decisions, these organizations will indirectly improve the customer experience through presumably higher returns.

Which goals describe why you are adopting AI at your organization?



Which of the following will your company use AI to address?

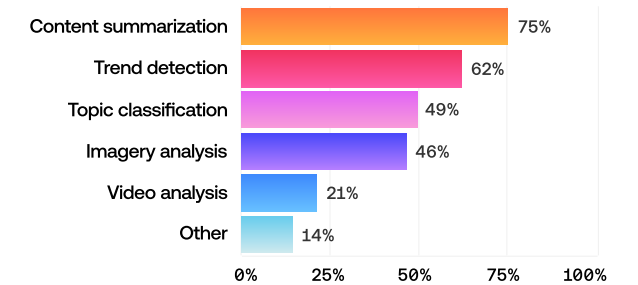


Financial Services

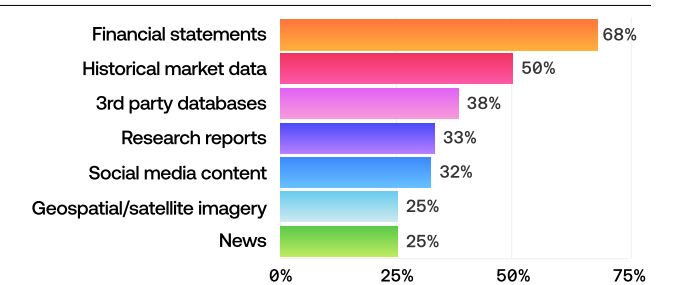
Content summarization includes summarizing data sources such as financial statements, historical data, news, and social media. Trend detection is applying AI to data to help identify patterns humans are otherwise ill-equipped to detect.

Financial services companies are using financial statements, historical market data, and 3rd party data in their investment models. Fewer are relying on social media content and geospatial/satellite imagery.

How will you leverage AI for investment research?



What data sources are used in your investment models?





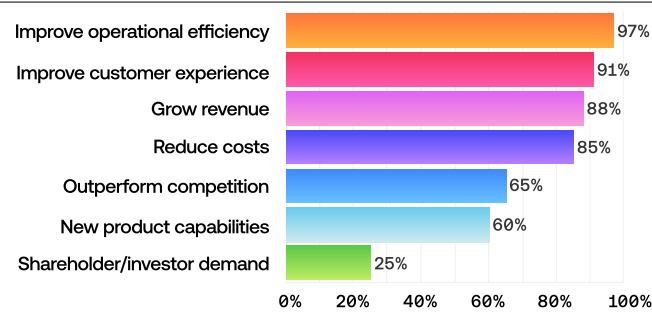
Logistics and Supply Chain

Logistics and supply chain companies adopt AI to help them improve operational efficiency, improve customer experience, and grow revenue.

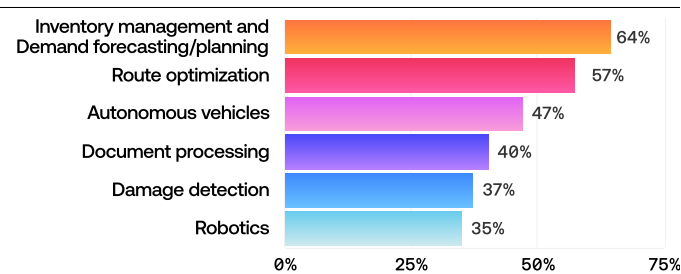
To help achieve these goals, logistics and supply chain companies are looking to adopt AI for better inventory management and demand forecasting, improved route optimization, to deploy autonomous vehicles, and improve document processing throughput and quality. These tools directly impact operational efficiency, which has downstream impacts on the overall customer experience, with reliable delivery and fewer delays.

For inventory management and demand forecasting, logistics and supply chain companies are adopting AI to help reduce costs, improve customer satisfaction, and improve forecast accuracy.

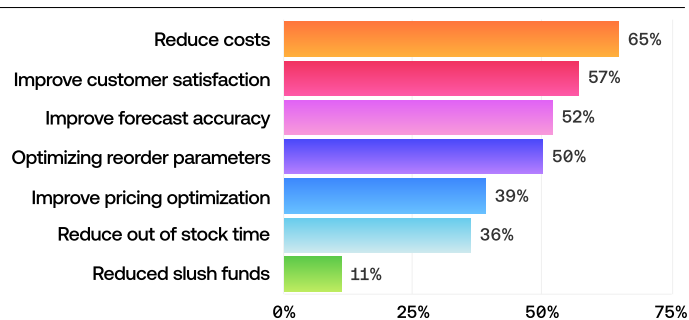
Which goals describe why you are adopting AI at your organization?



Which of the following will your company use AI to address?



What are your expected outcomes from implementing AI for Inventory management and demand forecasting/planning?



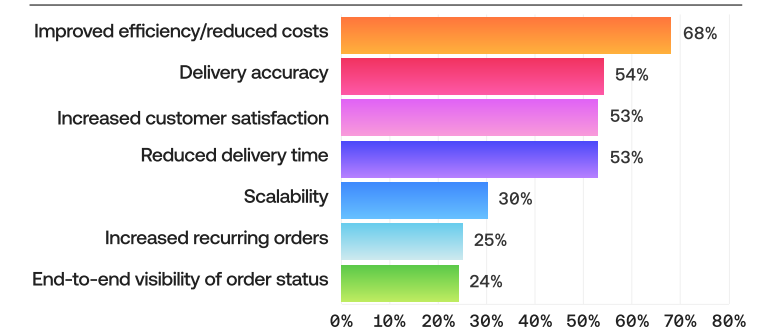
Logistics and Supply Chain

For route planning, logistics and supply chain companies believe AI can help improve efficiency, reduce costs, improve delivery accuracy, and reduce shipping times. This directly translates to improved operational efficiency and a better customer experience while indirectly translating to revenue growth.

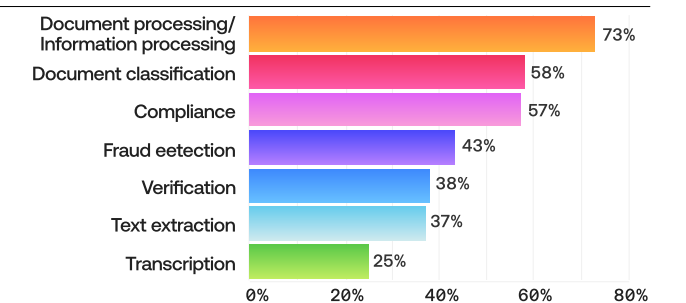
For document processing, logistics and supply chain companies look to AI to help them with information processing, document classification, and compliance. This application is strictly dedicated to increasing operational efficiency and reducing costs.

Logistics and supply chain companies process a lot of paperwork. To improve operational efficiency, this paperwork must be processed as quickly and accurately as possible. Logistics documents, such as bills of lading, commercial invoices, and packing lists are full of critical information required to clear shipments past customs and onto warehouses for distribution. Traditional OCR applications require the creation of templates for each type of document, which is infeasible and inefficient for global logistics companies. Instead, these companies rely on machine-learning-powered document processing, which requires no templates and still processes the documents at over over [95% accuracy](#).

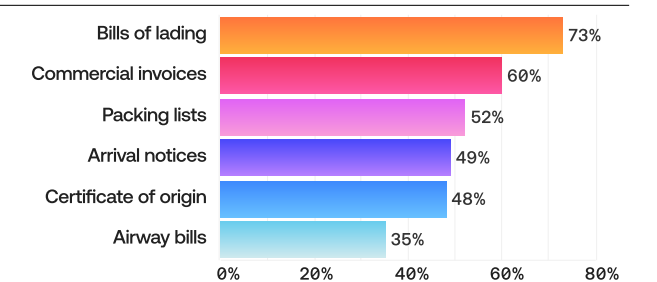
What outcomes are you expecting to achieve from implementing AI for route optimization?



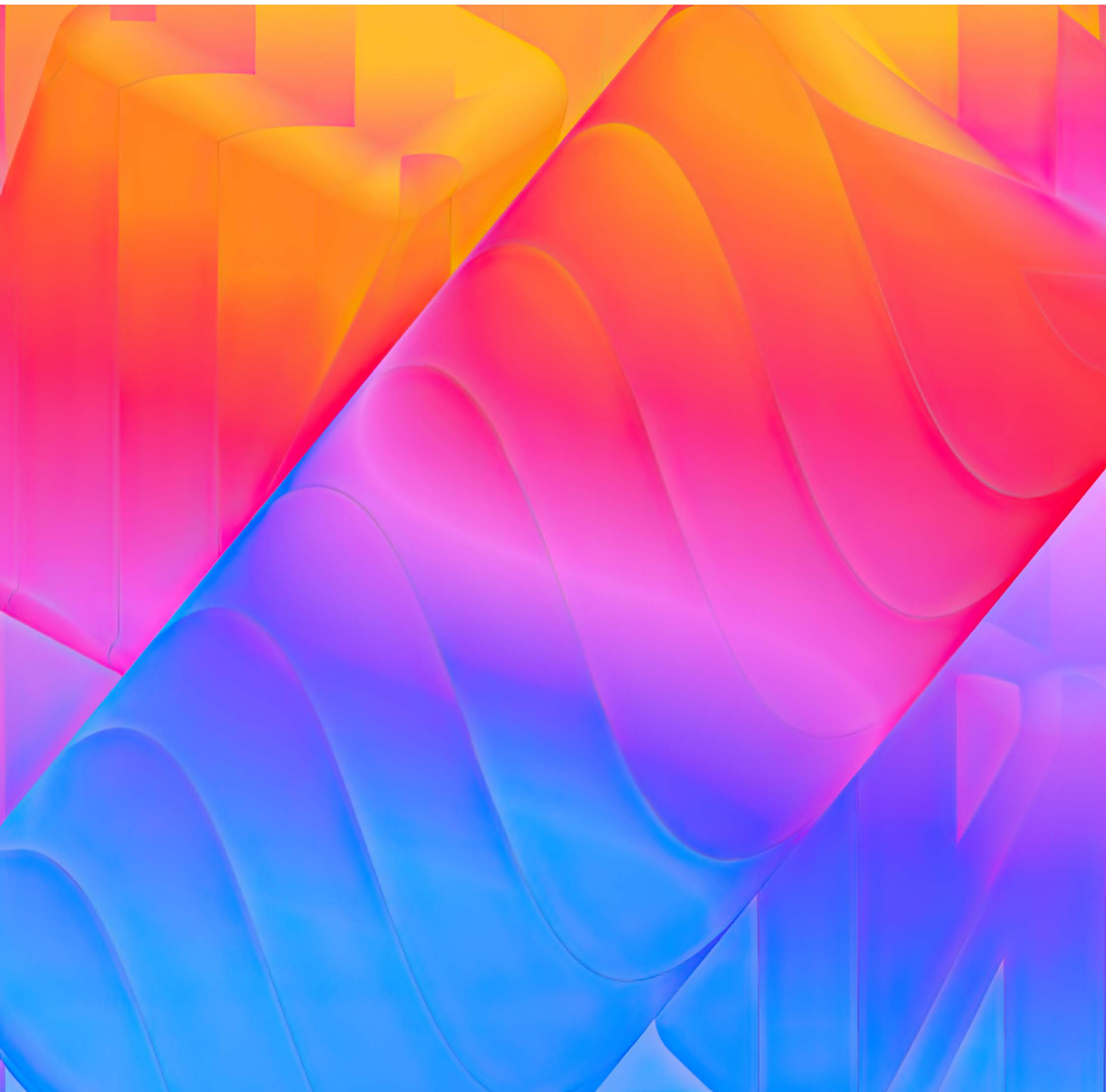
How will you use AI for document processing?



Which types of documents do you process?



ML Lifecycle

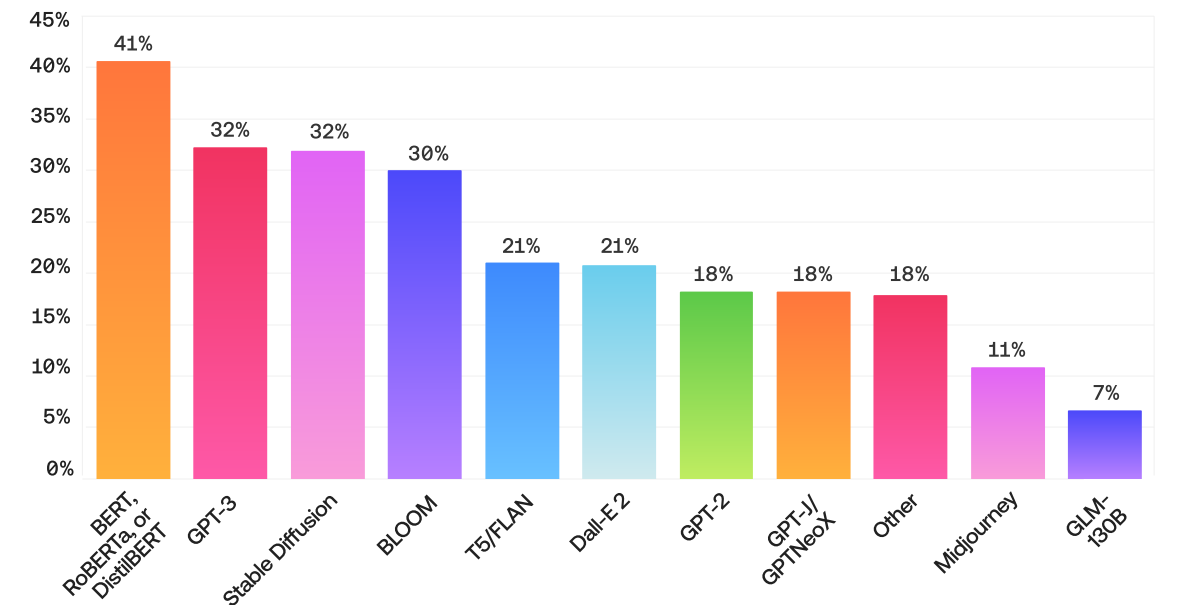


Working with Foundation Models

As of late 2022, the most widely used foundation models were BERT, GPT-3, Stable Diffusion, BLOOM, and T5/FLAN.

However, this landscape is quickly changing as more and more powerful generative models are being developed.*

Which generative models do you work with?



BERT plays a critical but quieter role in many organizations today, providing natural language understanding capabilities at a significantly reduced cost compared to larger models such as GPT-3.

However, this trend is shifting as BERT is not a generative model, so its use cases are limited compared

to state-of-the-art models like GPT-4. The compute required to run these more sophisticated models will become cheaper over time, and companies will have more third-party tools and expertise available to help integrate these larger models into their operations. Open-source models such as LLaMA are already being optimized to run on consumer laptops.

HELP IMPROVE THESE INSIGHTS

We'd appreciate your help to improve these insights! Given the rapid progress of Generative AI and the frequent introduction of new models, we're curious to learn which models are now the most widely adopted. Please take a moment to [fill out the survey](#).

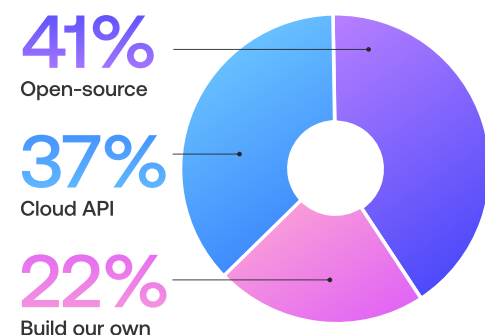
*(ChatGPT/GPT-3.5/GPT-4 were not included in this survey as survey collection began in late 2022 before these models were launched)

As previously discussed, of those companies that plan on working with generative models, the vast majority are looking to leverage open-source generative models (41%) or cloud API generative models (37%), while very few are looking to build their own generative models (22%).

The majority of companies using Cloud API LLMs are using OpenAI (64%), followed by AI21labs (26%) and Cohere (26%). Google and Anthropic recently launched their own LLMs, and several other companies expect to launch their own models in 2023.

To get the most value out of large generative models, many enterprises will need to fine-tune foundation models using their proprietary data and knowledge bases. The biggest challenges for fine-tuning foundation models are acquiring training data and the necessary ML infrastructure. Organizations working with generative models find it challenging and resource-intensive to fine-tune their models in-house. The most effective techniques, like RLHF, require humans to apply feedback to model outputs, custom software, and specialized skill sets to ensure high quality and limit human biases in models.

How do you work with generative models?



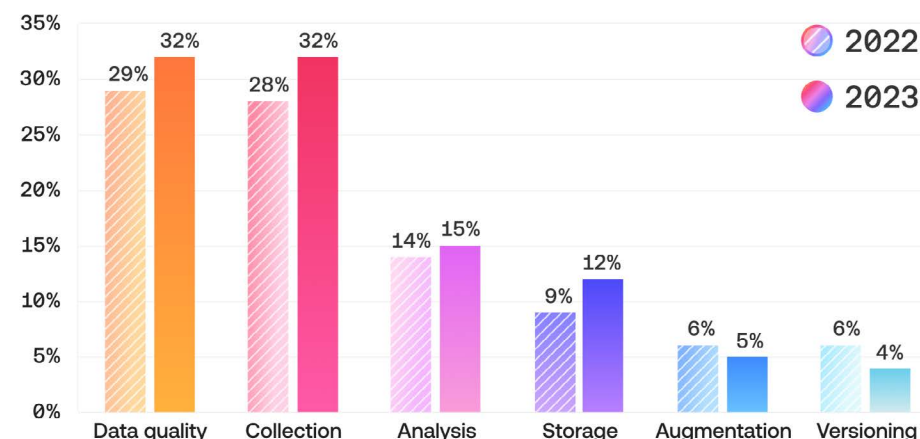
64% OpenAI

26% AI21labs

26% cohere

Data Challenges

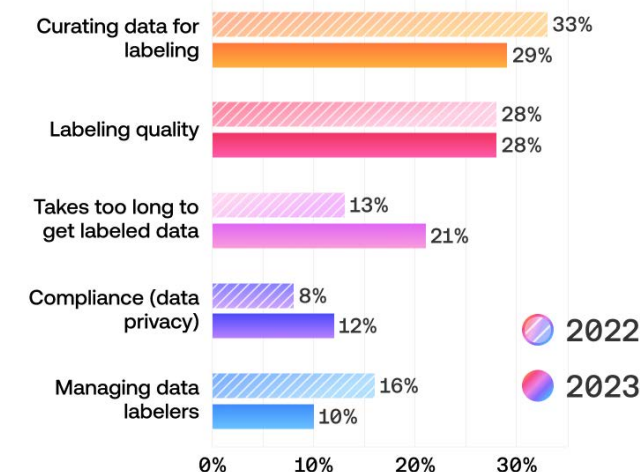
What are your team's biggest challenges with acquiring training data?



Data quality remains the most challenging part of acquiring training data, closely followed by data collection. The challenges for acquiring training data are nearly identical to last year's survey findings.

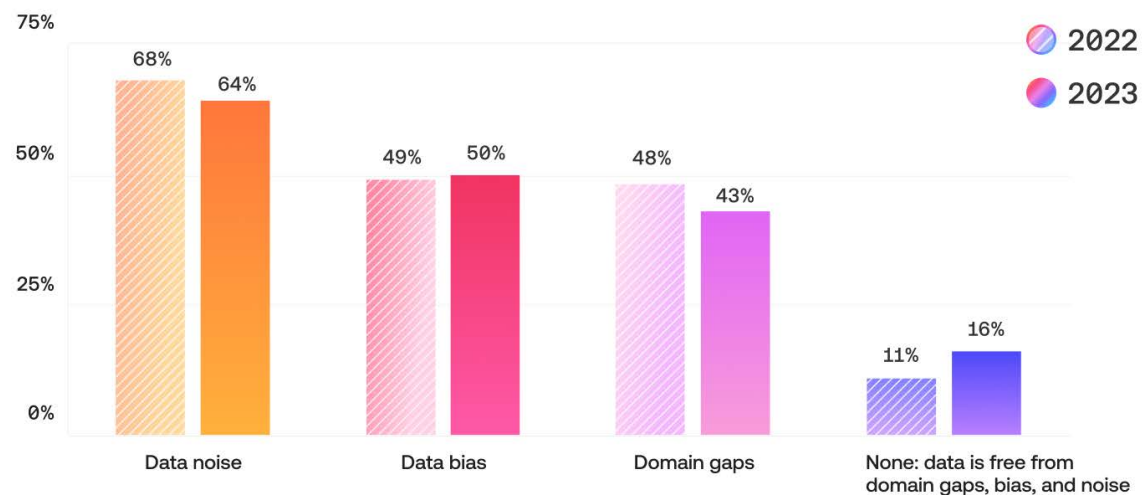
Curating data and annotation quality are still the top challenges for companies preparing training data for models. The biggest change YOY was that it takes too long to get labeled data (typically greater than one month), with 21% of respondents citing this as their #1 challenge, up from only 13% a year ago. Companies are looking to move faster to label their data, but it is difficult to keep up.

What are your biggest challenges with preparing data for training models?

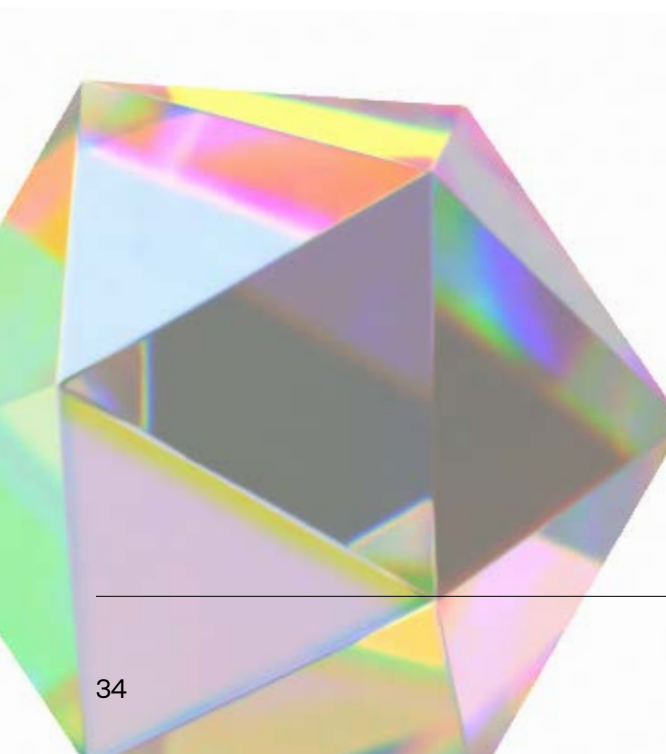


The majority of respondents continue to have problems with their training data.

What best describes the state of your training data?



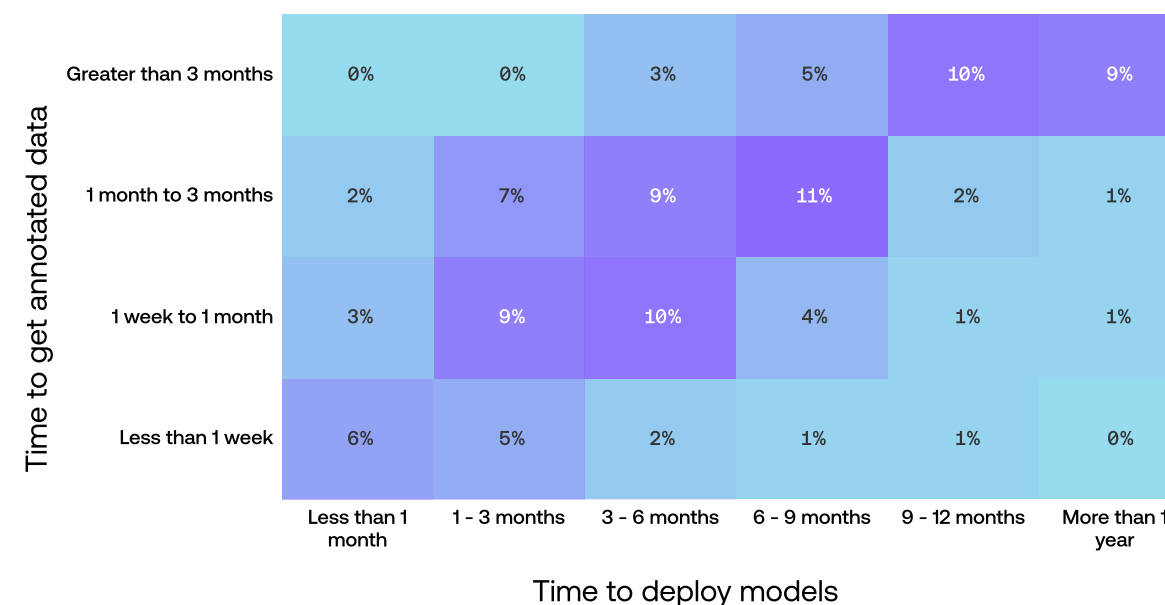
Most respondents continue to face problems with the quality of their training data. Data noise remains the number one issue (64%), followed by data bias (50%), and domain gaps (43%).



Data Best Practices

Companies that invest in good data annotation infrastructure can deploy new models, retrain existing ones, and deploy them into production faster.

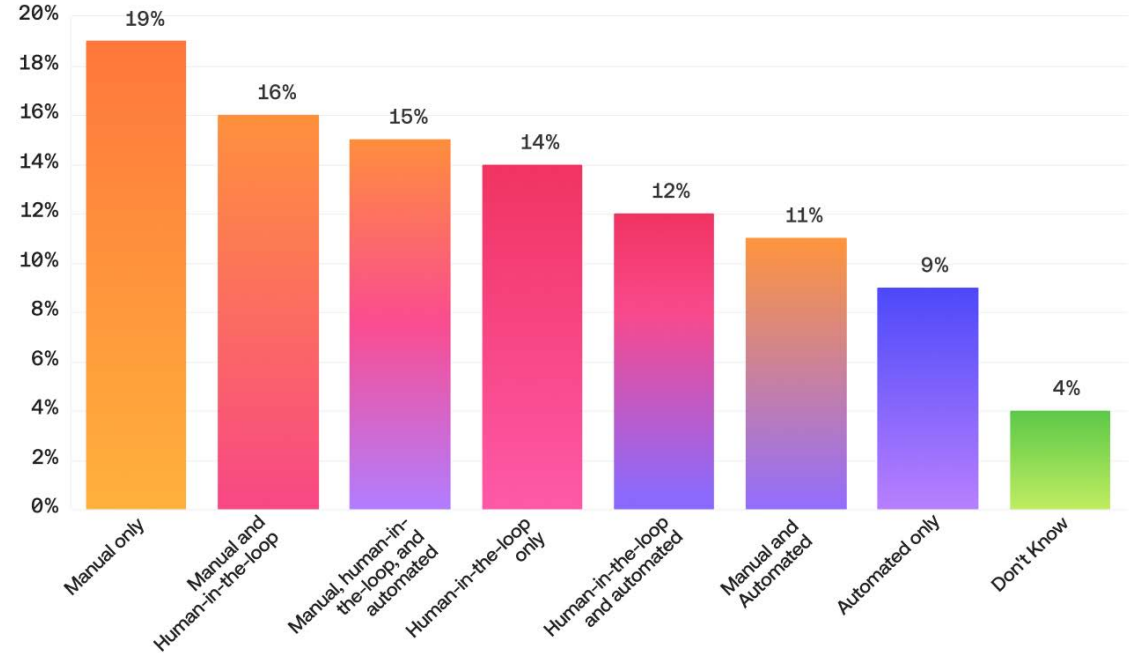
Teams that get annotated data faster tend to deploy new models to production faster and are able to update existing models more frequently.



Our data show that there seems to be a correlation between teams that get annotated data faster and teams that deploy new models to production faster and are able to update existing models more frequently.

Model Evaluation

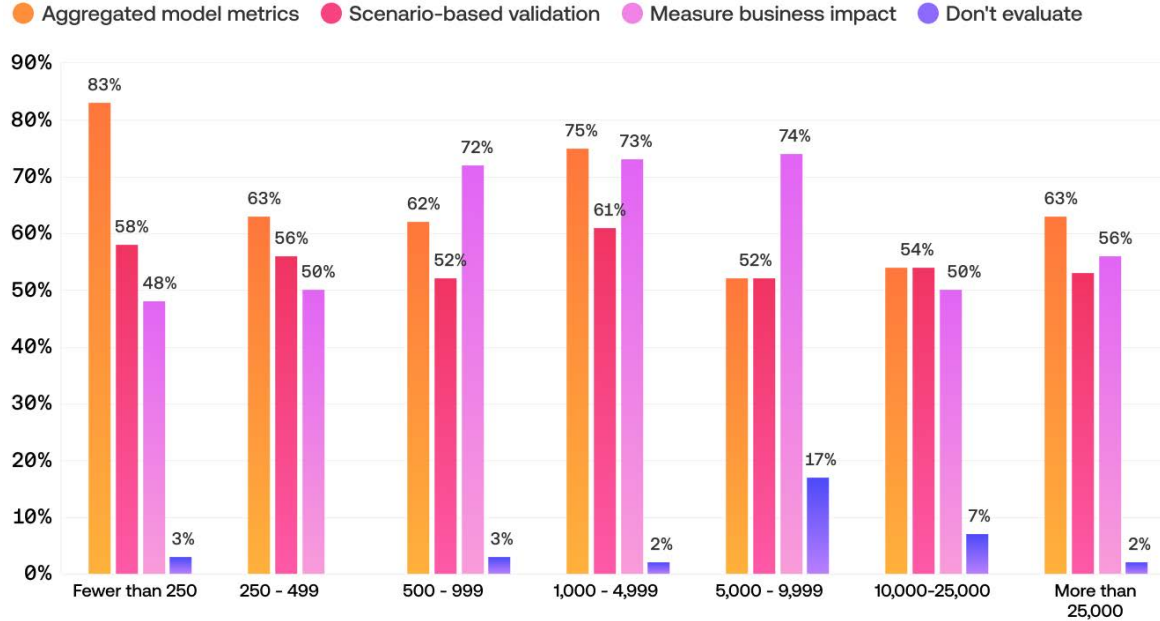
How are you labeling data when quality is the #1 challenge?



Respondents that only label their data manually most frequently rank labeling quality as their number one challenge in preparing training data, while those that leverage human-in-the-loop and automated labeling rank this challenge less frequently. 60% of respondents leveraging manual labeling in some capacity

rank labeling quality as their number one challenge, compared to 47% using automated labeling and only 44% using human-in-the-loop labeling. The combination of automated labeling plus [human-in-the-loop](#) is recommended as a best practice as it nearly always outpaces the accuracy and efficiency of either alone.

How do you evaluate model performance today?



While most of the report has focused on RLHF and Generative AI, companies apply machine learning in many different ways, from object detection to recommendation systems. One critical component of these production ML systems is how companies evaluate and monitor their performance.

Just as we found in 2022, measuring the business impact of models remains a challenge, especially for startups or very small companies (those with fewer than 250 employees). These companies rely more on aggregated model metrics as they are building products and a customer base, so the business impact is difficult to measure. However, small to medium size organizations (500-9,999 employees) are measuring business impact more than they were even one year ago, at about 73% now (compared to 55% in last year's survey).

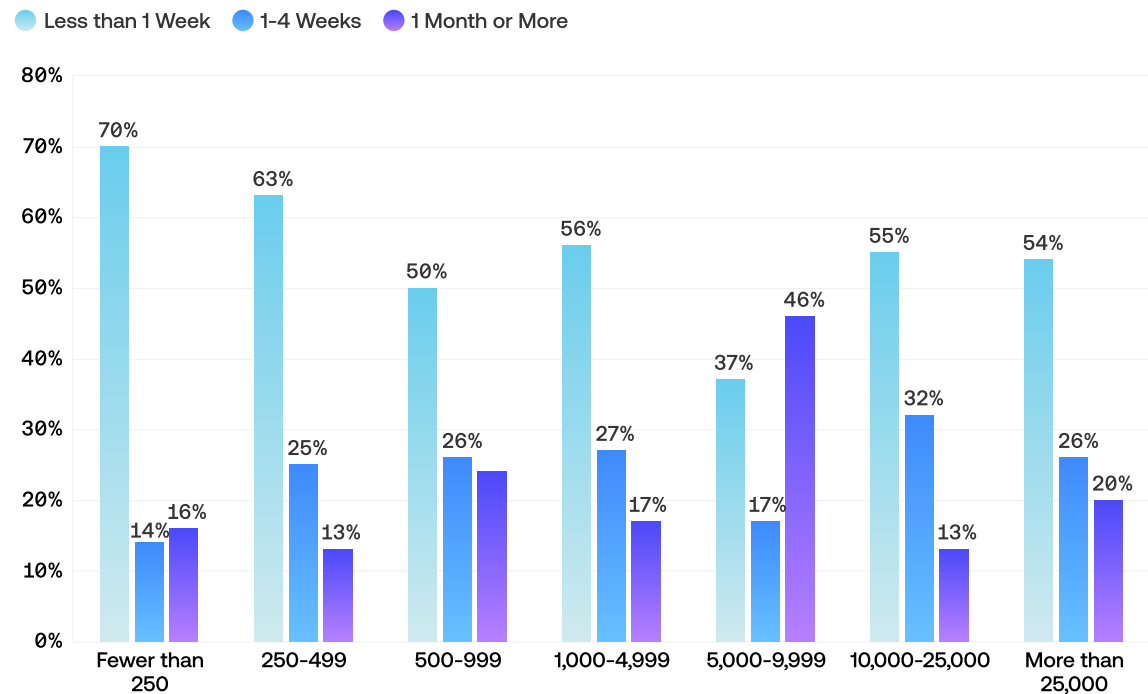
Larger companies take longer to identify issues in models

As in last year's report, smaller organizations can usually identify issues with their models quickly, in less than one week. Larger companies (those with 5,000 employees or more) may be more likely to measure business impact but are more likely than smaller ones to take longer (one month or more) to identify issues with their models.

These larger companies are operating complex systems at scale, and the issues they face are less likely to cause immediate business impact, so they are not

caught as quickly as startups or small companies, which are operating simpler systems and closely monitoring model metrics. Additionally, these larger companies typically have a larger customer base, so their users will hit on more edge cases than the smaller companies. For these reasons, it is important that as companies scale, they continually refine their MLOps practices, use [data curation](#) tools that can help them identify edge cases, and monitor their models while continuing to measure business impact.

How long does it typically take to discover an issue with the model?



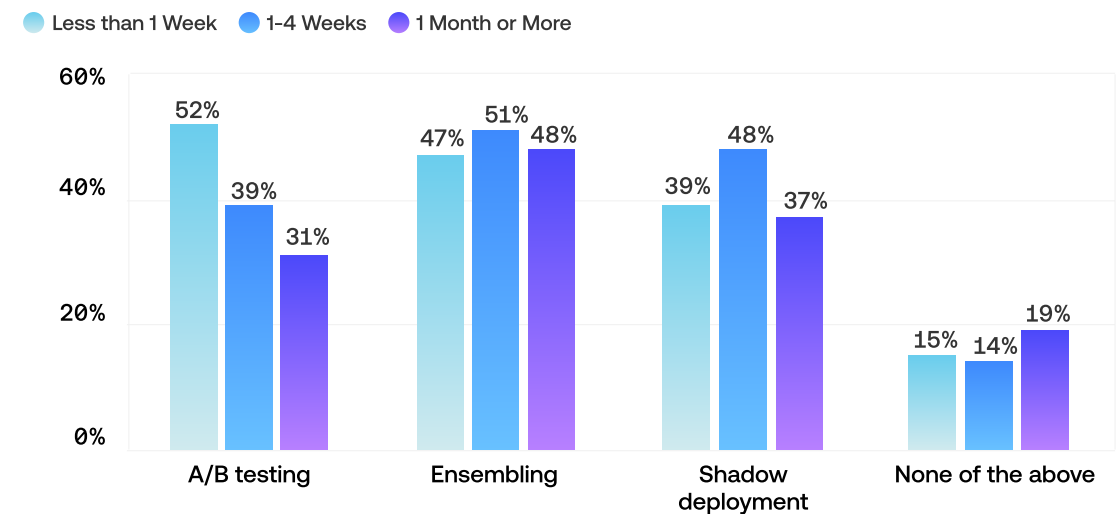
Model Evaluation Methods vs. Time to Deployment

As we found in last year's report, ML teams that identify issues fastest with their models are most likely to use A/B testing when deploying models.

Aggregate metrics are a useful baseline, but as enterprises develop a more robust modeling practice, tools such as "shadow deployments," ensembling, A/B testing, and even scenario tests can help validate models in challenging edge-case scenarios or rare classes.

Although small, agile teams at smaller companies may find failure modes, problems in their models, or problems in their data earlier than teams at large enterprises, their validation, testing, and deployment strategies are typically less sophisticated. Thus, with simpler models solving more uniform problems for customers and clients, it's easier to spot failures. When the system grows to include a large market or even a large engineering staff, complexity and technical debt begin to grow. At scale, scenario tests become essential, and even then, it may take longer to detect issues in a more complex system.

Model Deployment Method vs. time to discover model issues



While we have covered best practices for evaluating models, we must note that for the first time, the performance of Generative AI models is nearly impossible to evaluate automatically. This is because there are many ways for the model to respond

appropriately, and human judgment is required to evaluate correctness. That means any sound model evaluation, whether done by a generative model builder or an enterprise, will require human-in-the-loop validation and verification on an ongoing basis.

Conclusion

Generative AI is rapidly transforming the world, and businesses need to understand how to adopt this technology quickly or get left behind.

The most significant AI and ML readiness trend has been the enormous impact of Generative AI on companies, large and small, across all industries. While the 2022 AI Readiness Report focused on companies with in-house machine learning expertise, this year's report examined how all companies can adopt AI. This change in focus reflects the zeitgeist: dramatic improvements in the capabilities of Generative AI to accelerate innovation and transform every business.

We found that many companies plan to work with or experiment with foundation models, but many lack the expertise and tools to get these models into production. Most companies are adopting AI to enhance the customer experience, optimize operational efficiency, or improve profitability. Generative models will become increasingly more useful and widely accessible, making them essential to every organization's business strategy with a similar impact to the internet. Early adopters of AI are seeing the improved ability to develop new products or services, enhanced customer experience, and better collaboration across business functions, in addition to improved revenue and profitability.

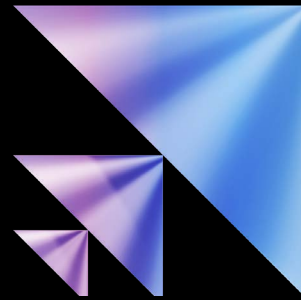
Enterprises widely use older non-generative models like BERT, but have realized they must adopt more generative models to stay ahead of the competition. Companies that fine-tune foundation models find their most significant challenges are acquiring training data, ML infrastructure, and comparing experiments across different models. Human evaluation has replaced benchmarks as the de-facto method to analyze large generative models and determine how they will work in a specific enterprise. Enterprises and governments need to leverage their unique data to unlock the full potential of generative models.

At Scale, our mission is to accelerate the development of AI applications to power the most ambitious AI projects in the world. To support that mission, we are excited to share the results of the Scale Zeitgeist: AI Readiness Report with you. We will continue to shed light on what it really takes to adopt AI and help you separate the signal from the noise.

“Large generative models are already giving people a productivity boost—we’ve seen how these systems help people write, code, learn, and more. We expect the capabilities of these models to rapidly improve, possibly beyond our imagination. If we can learn how to safely integrate AI into businesses by creating helpful, harmless, and honest systems, it could have a transformative effect on the economy and industries as we know them.”

**—JARED KAPLAN
CHIEF SCIENTIST, ANTHROPIC**

About Scale



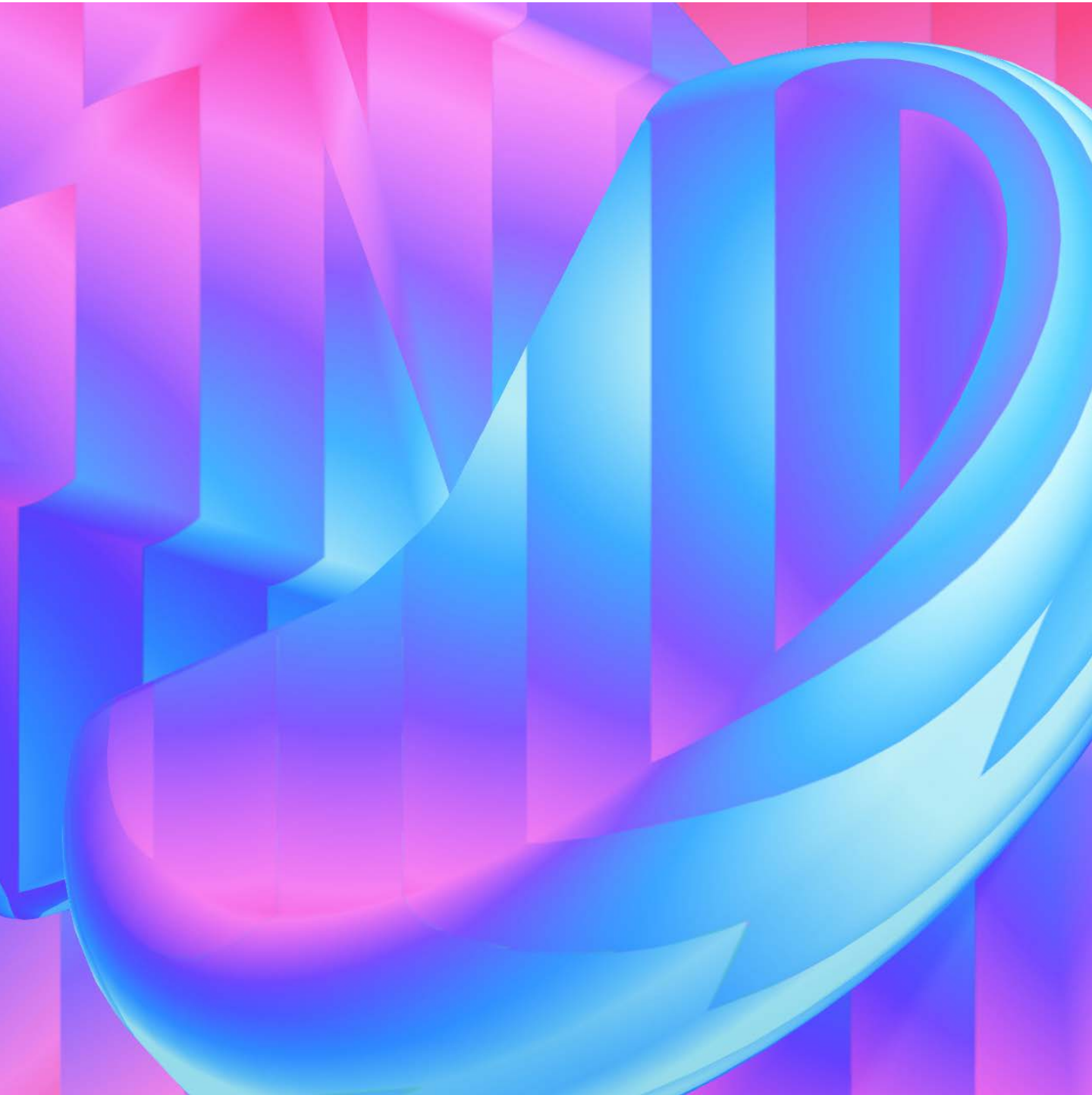
At Scale, our mission is to accelerate the development of AI applications. We believe that to make the best models, you need the best data.

The Scale Enterprise Generative AI Platform leverages your enterprise data to customize powerful base generative models to safely unlock the value of AI. The Scale Data Engine consists of all the tools and features you need to collect, curate and annotate high-quality data, in addition to robust tools to evaluate and optimize your models. Scale powers the most advanced LLMs and generative models in the world through world-class RLHF, data generation, model evaluation, safety, and alignment.

scale.com



Methodology



This survey was conducted online within the United States by Scale AI from December 15, 2022, to January 25, 2023. We received 2,909 responses from ML practitioners (e.g., ML engineers, data scientists, development operations, etc.) and leaders involved with AI in their companies. After data cleaning and filtering out those who indicated they are not involved with AI or ML projects and/or are not familiar with any steps of the ML development lifecycle, the dataset consisted of 1,699 respondents. We examined the data as follows: When asked to describe their level of seniority in their organizations, over one-third of respondents (37%) reported they are an individual contributor, nearly one quarter (25%) said they function as a team lead, 33% are a department head or executive, and 4% are owners. Most come from small companies with fewer than 500 employees (39%) or large companies with more than 10,000 employees (21%).

24% of respondents represent financial services/insurance, 22% represent the software/Internet/telecommunications industry, followed by retail and eCommerce (13%), logistics and supply chain (9%), education (7%), business and customer services (6%), automotive (5%), healthcare and life

sciences (4%), media/entertainment/hospitality (3%), manufacturing (2%), and other (6%).

Many respondents (31%) represent organizations that are advanced in their AI/ML adoption—they have multiple models deployed to production and are regularly retrained. About 18% are slightly less advanced—they have multiple models deployed to production—while 8% have only one model deployed to production, 12% are developing their first model, and 11% are only evaluating use cases.

